Some Aspects of Medical Research

THE GENETIC CODE

Two great families of molecules control the key functions of every living cell. They are the proteins and the nucleic acids. The main function of the proteins is to act as enzymes—the highly specialized catalysts which speed up chemical reactions in the cell, each acting in a specific way to promote a particular chemical reaction; under the mild conditions of temperature and acidity within a cell most of these reactions would only take place extremely slowly if enzymes were not present, and the cell could not function properly. The nucleic acids constitute the cell's genetic material—taking ' genetic ' in its general sense of relating to nuclear control of cellular function as well as to the transmission of ' blueprints ' from one generation to the next.

The structure of proteins

Protein molecules are large, typically containing thousands of atoms. Nevertheless, their basic chemical structure is remarkably simple. They consist of one or more polypeptide chains—that is, long chains with a backbone having a regular repeating structure to which different side-groups of atoms are attached at regular intervals. The structural units of this chain—the monomers which are joined together to form the protein polymer—are amino acids, of which there are 20 different kinds. A particular protein, which may be several hundred amino acid units long, has these amino acids arranged in a particular sequence. Since Sanger, of the Council's Laboratory of Molecular Biology in Cambridge, carried out his classic work on the structure of insulin, the amino acid sequences of several proteins have been experimentally determined.

This amino acid sequence is the so-called first-order structure of the protein. Once formed, the chain folds on itself in a precise but complicated way, so that each protein has an intricate three-dimensional shape peculiar to itself and different from that of all other proteins. It is this which allows each protein to carry out its special job. Thus proteins, as a family, are delicate, subtle and versatile. But behind this complexity their basic chemical structure—the linear sequence of amino acid units—is fairly uncomplicated, which means that they can be put together by a relatively simple process.

The structure of nucleic acids

It now seems likely that specifying the sequence of the 20 amino acids in each of the thousands of different proteins occurring in a living organism is the main function of the genetic material in the cell. This genetic material is contained in the genes, the units of genetic function, which are arranged in a linear order along the chromosomes. In higher organisms the chromosomes, with their complement of genes, reside in the nucleus of the cell. Each particular gene probably contains the instructions for making one particular protein: this is the 'one gene—one enzyme ' hypothesis.

It is now believed that genes are made from the other great family of biological molecules, the nucleic acids. There are two kinds of nucleic acid, closely related to each other, known as DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The genetic material is usually DNA, though in some small viruses, such as poliovirus, it is RNA; most of the RNA in cells has other, though related, functions.

DNA, like protein, is a polymer, and a very long one. When isolated from cells the molecules of DNA are usually at least 10 000 monomer units long, but they are easily broken and they may be longer than this inside the cells. The backbone of the DNA chain is made up of a regular, alternating sequence of two units: a phosphate group and a sugar group. Attached to each sugar group is a special side-group of atoms known as a base. Unlike protein, however, with its 20 side-groups, DNA commonly has only four different kinds—adenine, thymine, guanine and cytosine. These bases follow one another in an irregular order, and it is their precise order along the particular length of DNA constituting one gene which represents a genetic ' message '.

In fact DNA molecules usually consist of a pair of chains wound round each other into a helix, with the bases on each chain joining across the middle to form ' base pairs ', rather like the steps of a spiral staircase (Plate I). In this structure (which has been worked out by Crick, Watson and Wilkins), adenine on one chain pairs with thymine on the other, and similarly guanine pairs with cytosine. It is this double complementary structure which allows the cell to produce an exact copy of any DNA molecule when it divides and the strands separate, since only the correct sequence will ' fit '. A simple organism such as the bacteriophage *T4*, the virus that attacks the intestinal bacterium *Escherichia coli*, has about 200 000 base pairs. *E. coli* itself has perhaps 3 000 000, and man a few thousand million base pairs in each cell—enough for over a million genes if each gene is a few thousand base pairs long. When uncoiled, the DNA from all the cells in one human body would reach across the solar system.

In bacteriophage *T4* all the DNA is in the form of one molecule, some 60 μ in length. This constitutes the ' chromosome ' of the phage. It is likely that the chromosome of *E. coli* is also one single molecule, perhaps 1 mm long; but we do not know yet how the DNA is organized in the chromosomes of higher organisms. The phage chromosome contains about 100 genes, and these genes are ordered linearly on this structure. Each gene is perhaps 1000 base pairs long, and it is the order of the four kinds of base pairs along this length of DNA that determines in some way the precise order of the 20 amino acids of the protein specified by the gene. Working out how this is done has come to be known as the coding problem.

The coding problem

The basic difficulty in solving the code has been that, while in favourable circumstances the amino acid sequence of a protein can be determined, it is not yet technically possible to find the base sequence of a particular piece of DNA. The problem has thus to be attacked by indirect methods.

The first question we can ask is how many bases are needed to determine one amino acid. If only two of the four bases were used we should have only $4 \times 4 = 16$ possible combinations, whereas there are at least 20 kinds of amino acids in proteins. Thus the minimum number of bases needed is 3. There are $4 \times 4 \times 4 = 64$ possible triplets and it is not obvious how they should be allocated in relation to the 20 amino acids. For example, each amino acid might be made to the specification of just one triplet and the other 44 triplets might be ' nonsense '—that is, have some other function. Alternatively the code might be ' degenerate '—that is, several triplets might stand for each amino acid.

In either case one might expect one or more triplets to stand for a space between genes, or even that there would be separate triplets for ' begin chain ' and ' end chain '.

The earliest type of code suggested (Gamow, 1954) was an overlapping one. This is illustrated in Fig. 1, which shows how an overlapping triplet code would work. As can be seen, the first, second and third bases code the first amino acid, the second, third and fourth the second amino acid, and so on. It is easy to see that with such a code some sequences cannot be coded, and in fact it was soon

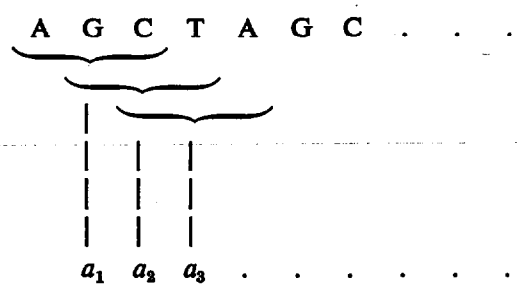

Figure 1

How an overlapping code would work (where A=adenine, G=guanine, C=cytosine, T=thymine, and a=amino acid).

shown that the actual family of codes proposed by Gamow must be incorrect, since they could not code some known sequences of amino acids. Later Brenner (1957), by an ingenious argument, showed that if the code was universal (the same in all living organisms), so that all the experimental data could be considered as a whole, then no simple overlapping triplet code was possible.

Recently more direct evidence has confirmed this (Tsugita, 1962; Wittmann, 1961). In an overlapping code a change of one base would, in general, alter *three* adjacent amino acids. Such changes, or mutations, may occur spontaneously in nature or they may be produced by chemical means, and all the data suggest that the typical alteration is to a *single* amino acid. It thus seems virtually certain that the code is not of the overlapping type.

If the code is not overlapping a new problem arises. How does one tell where one triplet ends and the next begins? For example, if the sequence in the middle of a gene is

. . . . CATCATCAT . . . . . .

is this to be read as

. . . . CAT CAT CAT . . . .
or . . . . C ATC ATC AT . . . .
or . . . . CA TCA TCA T . . . . ?

Various solutions to this problem have been suggested—see Crick, Griffith and Orgel (1957)—but it is now believed that none of these is correct. It seems likely that the message is read by starting from a fixed point and going along three at a time from there.

Experimental evidence

The evidence for this comes from genetic work carried out in the Council's Laboratory of Molecular Biology (Crick, Barnett, Brenner and Watts-Tobin, 1961). The system used was one of the two genes of the r_{II} locus of the virus that attacks *E. coli* (the *T*4 phage mentioned above). These are the genes so brilliantly exploited by Benzer in the United States—see below. The choice of these genes was dictated by the fact that the experiments can be done very quickly and that rare events can be studied: it is possible to handle very large numbers of the virus, and special techniques allow a required virus to be picked out from among a large number that are not required. For instance the type of phage *T*4 that contains the standard form of the genes—designated r_{II}^+—makes small turbid-edged plaques when grown in culture on strain *B* and strain *K* of *E. coli*. But a phage with altered (mutant) r_{II} genes makes a large clear-edged plaque on strain *B*, and does not grow at all on strain *K*. Hence the r_{II} mutants can be readily isolated when grown on strain *B*, and a small number of r_{II}^+ phages can be easily detected in any population of r_{II} mutants merely by ' plating ' on strain *K*.

The basic operation in ' genetic mapping ' experiments is the genetic cross. Two variations of the *T*4 phage are allowed to infect one bacterial cell at the same time. After 25 minutes the cell bursts open, and about 100 new viruses emerge. Some of these will be like the first ' parent ', and some like the second ' parent ', but in addition there will be some having mixed properties—that is, with some characteristics of the first parent and some of the second. Consider, for example, a case where one *T*4 phage has a defect at a point X on one of its r_{II} genes and another has a defect at a point Y on the same gene. When these are crossed together a few of the progeny will have defects at both X and at Y, and a few will have no defect. The nearer X and Y are together on the gene, the more rarely will recombination—that is, the presence or absence of both defects—occur, and if the mistakes are at the same point on the gene in the two phages, it is impossible to obtain a virus with neither defect. It was on the basis of this principle that Benzer (1959, 1961) was able to map the r_{II} genes and show that they had many different sites arranged in a linear array.

The genetic studies of Crick and his colleagues were made with a mutant gene of the *T*4 phage (one which had lost the capacity to allow the phage to grow on strain *K* of *E. coli*) produced by the action of proflavin, which, according to indirect evidence, brings about mutation by adding or deleting a base, rather than by changing one. When a stock of the altered virus was grown, occasionally a virus appeared in which the mutant gene was functioning again; this rare event (perhaps one in a million) could be picked out because of the powerful selective techniques that could be used to look for it. It was found, by the technique of genetic mapping, that this second alteration was not usually a correction at the original site of the defect, but was due to a further change at a nearby site. Either of these alterations, *by themselves*, could abolish the function of the gene —that is, the ability to grow on strain *K*—but when both alterations were together in the same gene the function was restored, though not completely.

This suggested the following explanation. If the genetic message is read in groups of three from one end, then the addition of a base near the beginning of the message will alter the reading of all the following triplets. This explains why such a gene has no function. However, let us suppose that the *second* defect is

due to the *removal* of a base. Then, although the few triplets between the two alterations will be changed, the rest of the message will be restored to its original meaning (Fig. 2). This explains why the gene works, and also why it is not exactly the same as the original version.

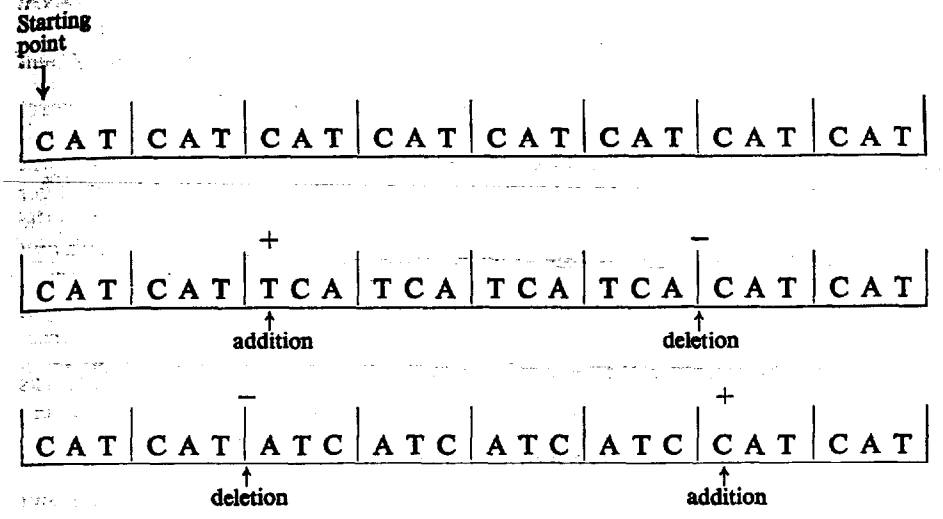

Figure 2

The effect of addition and deletion of bases on a triplet code. The letters C, A and T each represent a different base of the nucleic acid; for simplicity a repeating sequence is shown (this would code for a polypeptide chain in which only one kind of amino acid occurred).

It was possible to produce about 80 independent mutants within this region of the gene. They could be allocated to two classes, arbitrarily designated + and −, by seeing which pairs produced a workable gene. One can think of the ' + class ' as having an extra base, and the ' − class ' as having a base too few. By genetic methods, combinations of the type ' + with + ' and ' − with − ' were made and, as expected, these never had any function.

The most important discovery, however, was that three mutants of the *same* sign when combined into the same gene allowed the gene to function. This fits very well with the theory. Although the region between the added bases is altered, the rest of the gene is now restored to its original meaning. This result shows clearly that the ' coding ratio '—the number of bases which stands for one amino acid—is either 3 or a multiple of 3, depending on what we assume for the initial alterations. Subsidiary evidence suggests that the correct number is in fact 3, though 6 or 9 cannot be completely ruled out.

These results also suggest that the code is ' degenerate '—that there are not just 20 triplets that stand for amino acids and 44 that do not, but that in general an amino acid can be coded by any of several triplets; if it were not degenerate, combinations of the type ' + with − ' would not be likely to work when separated by the rather large distances observed. However, this is less certain than the rest of the conclusions.

Solving the code ' letters '

Although the genetic work shows the general nature of the genetic code, it would be impossible, or at least very difficult, to obtain the details of the code by genetic methods alone. This must be done by biochemical techniques.

It is believed that most protein synthesis in the cell takes place not on DNA, where the genetic material is stored, but in the cytoplasm. The actual sites are small, almost spherical, particles called ribosomes, which are roughly half RNA and half protein.

How is the genetic message transmitted to the ribosomes? At one time it was thought that the main RNA of the ribosomes was the ' messenger ', but recent work has suggested that a special RNA is formed, now called ' messenger RNA ', which is synthesized as a single-stranded copy of the DNA much in the same way that DNA copies itself (Jacob and Monod, 1961; Brenner, Jacob and Meselson, 1961; Gros *et al.*, 1961). This messenger goes into the ribosomes, where it acts as the actual template for protein synthesis; the classic slogan ' DNA makes RNA, and RNA makes protein ' thus seems to be essentially true.

The breakthrough in the biochemical approach to the coding problem was made by two scientists in America, Nirenberg and Matthaei (1961), who were trying to stimulate protein synthesis by adding virus RNA to a special cell-free system obtained from broken bacterial cells. They decided to try adding an artificial RNA, which had been synthesized by a simple enzyme system and which contained not all four bases but just one of them, uracil, repeated many times (uracil occurs in RNA instead of the related thymine in DNA). This material, known as poly U, when added to the cell-free system stimulated the production of polypeptide chains consisting only of the amino acid phenylalanine. Thus the RNA code for phenylalanine is probably the triplet of bases UUU.

This result was reported at the International Biochemical Congress at Moscow in August 1961, and it was immediately apparent that there was a very good chance that the code could be solved by further work along these lines. Although it is not yet possible to produce a long RNA molecule with a defined base sequence, very short molecules, having up to three or four bases with defined sequences, can be obtained, and these short molecules can be used to start the enzymatic synthesis of long chains. In addition, polymers of known composition but of random sequence can be produced—for example, poly UC, made of uracil and cytosine arranged at random. It has been found in several laboratories that if this is added to the cell-free system the polypeptide chains produced contain only four of the twenty amino acids, namely phenylalanine, serine, leucine and proline. Such work is now in full flood, and the results show without doubt that the artificial RNAs are fairly specific in the effects they produce (Matthaei *et al.*, 1962; Lengyel *et al.*, 1962).

However, even if this approach is completely successful, so that we can say in detail which triplets correspond to which amino acids, two other aspects of the problem remain. It has still not been shown that there is a simple linear relationship between the gene and the amino acid sequence of the protein it produces. It would be surprising, however, if there were not, and there is now a reasonable hope that proof may be obtained for one or two cases within the next year or so.

The answer to the other problem is less certain. Is the code universal? We know that the same set of 20 amino acids is used throughout nature, from viruses to man, but are they always coded by the same triplets? There are several experiments which suggest that they are, for the most part at least. However, by the use of cell-free systems produced from different organisms the answers could easily be found once the synthetic RNAs are available.

Benzer, S. (1959). *Proc. nat. Acad. Sci. (Wash.)*, **45**, 1607.

—— (1961). *Proc. nat. Acad. Sci. (Wash.)*, **47**, 403.

Brenner, S. (1957). *Proc. nat. Acad. Sci. (Wash.)*, **43**, 687.

—— Jacob, F. and Meselson, M. (1961). *Nature (Lond.)*, **190**, 576.

Crick, F. H. C., Barnett, L., Brenner, S. and Watts-Tobin, R. J. (1961). *Nature (Lond.)*, **192**, 1227.

—— Griffith, J. S. and Orgel, L. E. (1957). *Proc. nat. Acad. Sci. (Wash.)*, **43**, 416.

Gamow, G. (1954). *Nature (Lond.)*, **173**, 318

Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Risebrough, R. W. and Watson, J. D. (1961). *Nature (Lond.)*, **190**, 581.

Jacob, F. and Monod, J. (1961). *J. molec. Biol.*, **3**, 318.

Lengyel, P., Speyer, J. F., Basilio, C. and Ochoa, S. (1962). *Proc. nat. Acad. Sci. (Wash.)*, **48**, 282.

Matthaei, J. H., Jones, O. W., Marton, R. G. and Nirenberg, M. W. (1962). *Proc. nat. Acad. Sci. (Wash.)*, **48**, 666.

Nirenberg, M. W. and Matthaei, J. H. (1961). *Proc. nat. Acad. Sci. (Wash.)*, **47**, 1588.

Tsugita, A. (1962). *J. molec. Biol.*, **5**, 284.

Wittmann, H. G. (1961). *Naturwissenschaften*, **48**, 729.